## **Dataset Description**

The WiDS Datathon++ 2024 focuses on a prediction task involving roughly 39k (split into training and test sets) of observations of Time to Treatment for Metastatic Breast Cancer. The dataset consists of patient characteristics (age, race, bmi, zip code), treatment characteristics (breast cancer diagnosis code, metastatic cancer diagnosis code, metastatic cancer treatments etc.) as well as social and economic predictors (income, education, rent, race, poverty etc associated with the patient zip code.) Each row in the data corresponds to the treatment\_period for a single patient. Your task is to predict the treatment\_period for each row, given the characteristics of the patient, treatment and related social and economic predictors.

You are provided with two datasets:  
**(1)** the training dataset train.csv where the observed values of the Treatment Period for each row is provided  
**(2)** the test dataset test.csv where we withhold the observed values of the Treatment Period for each row

To participate in the Datathon++, you will submit a solution file containing the predicted Treatment Period values for each row in the test dataset. The predicted values you submit will be compared against the observed Treatment Period values for the test dataset and this will determine your standing on the Leaderboard during the competition as well as your final standing when the competition closes.

You are also provided with an example of a solution file prepared for submission.

## **External Data Usage**

The datathon task can be tackled successfully without the use of external data. In fact, the degree to which we have anonymized the data would make joining additional data to the competition data difficult. However, participants who wish to do so may use additional external data for the purpose of building predictive models.

## **Data Dictionary**

**Covariates**For this year’s Datathon++, we are using a real world evidence dataset from Health Verity, one of the largest healthcare data ecosystems in the US, as the main data source for the Datathon. In particular, the HV dataset that we use for the Advance challenge contains health related information of patients who were diagnosed with metastatic triple negative breast cancers from the year 2015- 2018 in the US. We also enriched the data set with the US Zip Codes Database which were built from the ground up using authoritative sources including the U.S. Postal Service™, U.S. Census Bureau, National Weather Service, American Community Survey, and the IRS, to obtain additional social economic information based on the locations of the patients.

**Target**

* treatment\_pd: This is the difference (in days) between metastatic first treatment date and breast cancer diagnosis date.

**Columns**

* patient\_id - Unique identification number of patient
* patient\_race - Asian, African American, Hispanic or Latino, White, Other Race
* payer\_type - payer type at Medicaid, Commercial, Medicare on the metastatic date
* patient\_state - Patient State (e.g. AL, AK, AZ, AR, CA, CO etc…) on the metastatic date
* patient\_zip3 - Patient Zip3 (e.g. 190) on the metastatic date
* patient\_age - Derived from Patient Year of Birth (index year minus year of birth)
* patient\_gender - F, M on the metastatic date
* bmi - If Available, will show available BMI information (Earliest BMI recording post metastatic date)
* breast\_cancer\_diagnosis\_code - ICD10 or ICD9 diagnoses code
* breast\_cancer\_diagnosis\_desc - ICD10 or ICD9 code description
* breast\_cancer\_diagnosis\_year - Calendar Year of first breast cancer diagnosis
* metastatic\_cancer\_diagnosis\_code - ICD10 diagnoses code
* metastatic\_first\_treatment - Generic drug name of the first treatment (e.g. "Cisplatin") after metastatic diagnosis
* metastatic\_first\_treatment\_type - Description of Treatment (e.g. Antineoplastic) of first treatment after metastatic diagnosis
* metastatic\_first\_novel\_treatment - Generic drug name of the first novel treatment (e.g. "Cisplatin") after metastatic diagnosis
* metastatic\_first\_novel\_treatment\_type - Description of Treatment (e.g. Antineoplastic) of first novel treatment after metastatic diagnosis
* region - Region of patient location
* division - Division of patient location
* population - An estimate of the zip code's population.
* density - The estimated population per square kilometer.
* age\_median - The median age of residents in the zip code.
* male - The percentage of residents who report being male (e.g. 55.1).
* female - The percentage of residents who report being female (e.g. 44.9).
* married - The percentage of residents who report being married (e.g. 44.9).
* family\_size - The average size of resident families (e.g. 3.22).
* income\_household\_median - Median household income in USD.
* income\_household\_six\_figure - Percentage of households that earn at least $100,000 (e.g. 25.3)
* home\_ownership - Percentage of households that own (rather than rent) their residence.
* housing\_units - The number of housing units (or households) in the zip code.
* home\_value - The median value of homes that are owned by residents.
* rent\_median - The median rent paid by renters.
* education\_college\_or\_above - The percentage of residents with at least a 4-year degree.
* labor\_force\_participation - The percentage of residents 16 and older in the labor force.
* unemployment\_rate - The percentage of residents unemployed.
* race\_white - The percentage of residents who report their race White.
* race\_black - The percentage of residents who report their race as Black or African American.
* race\_asian - The percentage of residents who report their race as Asian.
* race\_native - The percentage of residents who report their race as American Indian and Alaska Native.
* race\_pacific - The percentage of residents who report their race as Native Hawaiian and Other Pacific Islander.
* race\_other - The percentage of residents who report their race as Some other race.
* race\_multiple - The percentage of residents who report their race as Two or more races.
* hispanic - The percentage of residents who report being Hispanic. Note: Hispanic is considered to be an ethnicity and not a race.
* age\_under\_10 - The percentage of residents aged 0-9.
* age\_10\_to\_19 - The percentage of residents aged 10-19.
* age\_20s - The percentage of residents aged 20-29.
* age\_30s - The percentage of residents aged 30-39.
* age\_40s - The percentage of residents aged 40-49.
* age\_50s - The percentage of residents aged 50-59.
* age\_60s - The percentage of residents aged 60-69.
* age\_70s - The percentage of residents aged 70-79.
* age\_over\_80 - The percentage of residents aged over 80.
* divorced - The percentage of residents divorced.
* never\_married - The percentage of residents never married.
* widowed - The percentage of residents never widowed.
* family\_dual\_income - The percentage of families with dual income earners.
* income\_household\_under\_5 - The percentage of households with income under $5,000.
* income\_household\_5\_to\_10 - The percentage of households with income from $5,000-$10,000.
* income\_household\_10\_to\_15 - The percentage of households with income from $10,000-$15,000.
* income\_household\_15\_to\_20 - The percentage of households with income from $15,000-$20,000.
* income\_household\_20\_to\_25 - The percentage of households with income from $20,000-$25,000.
* income\_household\_25\_to\_35 - The percentage of households with income from $25,000-$35,000.
* income\_household\_35\_to\_50 - The percentage of households with income from $35,000-$50,000.
* income\_household\_50\_to\_75 - The percentage of households with income from $50,000-$75,000.
* income\_household\_75\_to\_100 - The percentage of households with income from $75,000-$100,000.
* income\_household\_100\_to\_150 - The percentage of households with income from $100,000-$150,000.
* income\_household\_150\_over - The percentage of households with income over $150,000.
* income\_individual\_median - The median income of individuals in the zip code.
* poverty - The median value of owner occupied homes.
* rent\_burden - The median rent as a percentage of the median renter's household income.
* education\_less\_highschool - The percentage of residents with less than a high school education.
* education\_highschool - The percentage of residents with a high school diploma but no more.
* education\_some\_college - The percentage of residents with some college but no more.
* education\_bachelors - The percentage of residents with a bachelor's degree (or equivalent) but no more.
* education\_graduate - The percentage of residents with a graduate degree.
* education\_stem\_degree - The percentage of college graduates with a Bachelor's degree or higher in a Science and Engineering (or related) field.
* self\_employed - The percentage of households reporting self-employment income on their 2016 IRS tax return.
* farmer - The percentage of households reporting farm income on their 2016 IRS tax return.
* disabled - The percentage of residents who report a disability.
* limited\_english - The percentage of residents who only speak limited English.
* commute\_time - The median commute time of resident workers in minutes.
* health\_uninsured - The percentage of residents who report not having health insurance.
* veteran - The percentage of residents who are veterans.